Informative Sample Mining Network for Multi-Domain Image-to-Image Translation

Jie Cao[1,3][0000−0001−6368−4495], Huaibo Huang[1,3][0000−0001−5866−2283], Yi Li[1,3][0000−0002−2856−7290], Ran He[1,2,3 *][0000−0002−3807−991X], and Zhenan Sun[1,2,3][0000−0003−4029−9935]

[1] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[2] Center for Excellence in Brain Science and Intelligence Technology, CAS
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
{jie.cao, huaibo.huang, yi.li}@cripac.ia.ac.cn, {rhe, znsun}@nlpr.ia.ac.cn

Abstract. The performance of multi-domain image-to-image translation has been significantly improved by recent progress in deep generative models. Existing approaches can use a unified model to achieve translations between all the visual domains. However, their outcomes are far from satisfying when there are large domain variations. In this paper, we reveal that improving the sample selection strategy is an effective solution. To select informative samples, we dynamically estimate sample importance during the training of Generative Adversarial Networks, presenting Informative Sample Mining Network. We theoretically analyze the relationship between the sample importance and the prediction of the global optimal discriminator. Then a practical importance estimation function for general conditions is derived. Furthermore, we propose a novel multi-stage sample training scheme to reduce sample hardness while preserving sample informativeness. Extensive experiments on a wide range of specific image-to-image translation tasks are conducted, and the results demonstrate our superiority over current state-of-the-art methods.

Keywords: image-to-image translation, multi-domain image generation, generative adversarial networks

1 Introduction

Multi-domain image-to-image translation (I2I) aims at learning the mappings between visual domains. These domains can be instantiated by a set of attributes, each of which represents a meaningful visual property. Since each possible combination of the attributes specifies a unique domain [28], the total domain number can be huge in practical applications. For instance, if there are 10 independent binary attributes, we need to handle the translations among 1024 visual domains, which is far more challenging than the two-domain translation. Fortunately, thanks to the recent advances in deep generative models [7,18], current methods [2,11,21,36] can achieve multi-domain I2I by a single model. However,

Fig. 1: Facial attribute transfer results of different methods. From top to bottom, the target attributes of each row are "*beard*, *no_smiling*, *black_hair*", "*bangs*, *eyeglasses*, *black_hair*", and "*female*, *smiling*, *brown_hair*", respectively. While AttGAN [11] and STGAN [21] show degraded results in these challenging cases, our approach achieves effective attribute transfer and maintains the realism of texture details.

these methods only produce promising results when translating images within similar visual domains, e.g., changing human hair color. The translations between domains with large semantic discrepancies are still not well addressed yet.

To further illustrate the limitation of existing approaches, we take the task of facial attribute transfer as a prime example. Fig. 1 shows the results of some challenging translations. Due to the large gap between the source and the target domain, it is difficult to transfer target attributes without impairing visual realism. Even the current state-of-the-art methods [11,21] produce degraded results, although they can address most of the easy translations (we will show these cases in the following experiments). This phenomenon indicates that existing methods mainly focus on easy cases during training but neglect hard ones.

We argue that effective sample selection strategies greatly help to address this problem. Sample selection is within the scope of deep metric learning, which contributes to many computer vision tasks. The studies in deep metric learning [14,29,38,40] point out that a large fraction of training samples may satisfy the loss constraints, providing no progress for model learning. That is, the vast majority of samples are too easy, so their contributions to our training are only

marginal. Unfortunately, current multi-domain I2I methods merely adopt the naive random sample selection that treats all training samples equally and thus selects the easy ones mostly. It intuitively hinders training efficiency and consequently leads to the degradation discussed above.

In this paper, we propose **In**formative sample m**i**ning ne**t**work (INIT) to enhance training efficiency and improve performance in multi-domain I2I tasks. Concretely, we integrate Importance Sampling into the generation framework under Generative Adversarial Networks (GAN). Adversarial Importance Weighting is proposed to select informative samples and assign them greater weight. We derive the weighting function based on the assumption that the global optimal discriminator is known. Then we consider more general conditions and introduce the guidance from the prior model to rescale the importance weight. Furthermore, we propose Multi-hop Sample Training to avoid the potential problems [29,35,40] in model training caused by sample mining. Based on the principle of divide-and-conquer, we produce target images by multiple hops, which means the image translation is decomposed into several separated steps. On the one hand, our training scheme preserves sample informativeness. On the other hand, step-by-step training ensures that the generator can learn complex translations. Combining with Adversarial Importance Weighting and Multi-hop Sample Training, our approach can probe and then fully utilize informative training samples.

To verify the effectiveness of our approach, we conduct experiments on facial attribute transfer, season transfer, and edge&photo transfer. We make extensive comparisons with current state-of-the-art multi-domain I2I methods. The experimental results demonstrate our improvements in both attribute transfer and content preservation.

Our contributions can be summarized as follows:

- We analyze the importance of sample selection in image-to-image translation and propose Informative Sample Mining Network.
- We propose Adversarial Importance Weighting, which integrates Importance Sampling into GAN, to achieve effective training sample mining.
- We propose Multi-hop Sample Training to reduce the hardness of the probed informative samples, making them easy to train.
- We provide extensive experimental results on facial attribute transfer, season transfer, and edge&photo transfer, showing our superiority over existing approaches.

2 Related work

Image-to-Image Translation. Recent advances in deep generative models [7,18,26] have brought much progress in the field of image-to-image translation [2,3,11,15,19,21,22,36,41]. At the early stage, the studies are focused on translations between two visual domains. Zhu et al. have done pioneering works on learning the translations with paired data [15] and unpaired data [41]. FaderNet [19] disentangles the salient information in the latent space to control attribute intensity. UNIT [22] combines Variational AutoEncoder [18] with GAN,

and present high-quality results on unsupervised translation tasks. Later on, a lot of efforts are made to deal with the multi-domain condition. StarGAN [2] is the first unified model that produces visually plausible multi-domain translation results. AttGAN [11] introduces an attribute-aware constraint as well as a reconstruction-based regularization to achieve "only change what you want". Following AttGAN, Liu et al. [21] propose a novel selective transfer unit to enhance image quality. RelGAN [36] introduces the notion of relative attribute and employs multiple discriminators to improve both attribute translation and interpolation. Different from previous approaches, we make improvements from a new perspective: we probe informative samples during training, making our network aware of the most challenging cases.

Deep Metric Learning. Deep metric learning aims at learning good representations. The core idea is to narrow the distances of similar images in the embedding space and enlarge the distances of dissimilar ones. Existing works mainly focus on how to choose proper loss functions and sample selection strategies. In the studies of loss function, contrastive loss [13] and triplet loss [34] are the most representative works. The two losses are widely adopted and extended by successive methods. Huang et al. [14] explore the structure of quadruplets. Wang et al. [33] improve triplet loss by introducing a third-order geometry relationship. Sample selection strategies have also been widely studied. For example, hard negative sample mining [30] is proposed to replace the random sample selection in the contrastive loss. FaceNet [29] first adopts semi-hard negative mining within a batch for face recognition. Harwood et al. [9] utilize approximate nearest neighbor search to select harder samples adaptively. Recently proposed methods [4,39] introduce adversarial learning to generate potentially informative samples to train the model. At present, sample selection strategies have been adopted in many computer vision tasks, including image classification [20], face recognition [1,8,13,29], and person re-identification [37]. In this work, we show that sample strategy is also important to image generation but rarely studied. To this end, we propose a novel sampling strategy specifically for multi-domain I2I.

Importance Sampling. Importance Sampling (IS) [6] is a method for estimating properties of a target distribution $p_{data}(\boldsymbol{x})$ which is difficult to sample from directly. Samples are instead drawn from a proposal distribution $p_g(\boldsymbol{x})$ that over-weights the important region. IS is essential for many statistical theories, including Bayesian inference [6] and sequential Monte Carlo methods [25,32]. Applying IS, we can draw samples from $p_g(\boldsymbol{x})$ to estimate $\mathbb{E}_{p_{data}}[\mathcal{L}(\boldsymbol{X})]$ for any known function $\mathcal{L}$. Specifically, we have

$$\mathbb{E}_{p_{data}}[\mathcal{L}(\boldsymbol{X})] = \int \frac{p_{data}(\boldsymbol{x})}{p_g(\boldsymbol{x})}\mathcal{L}(\boldsymbol{x})p_g(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_{p_g}\left[\frac{p_{data}(\boldsymbol{X})}{p_g(\boldsymbol{X})}\mathcal{L}(\boldsymbol{X})\right], \quad (1)$$

where for any $\boldsymbol{x}$ in the sample space, we have $p_g(\boldsymbol{x}) > 0$ whenever $p_g(\boldsymbol{x}) \cdot p_{data}(\boldsymbol{x}) \neq 0$. The likelihood ratio $\frac{p_{data}(\boldsymbol{x})}{p_g(\boldsymbol{x})}$ is also referred to as the importance weight. In the following section, we will integrate IS into the GAN-based multi-domain I2I methods.

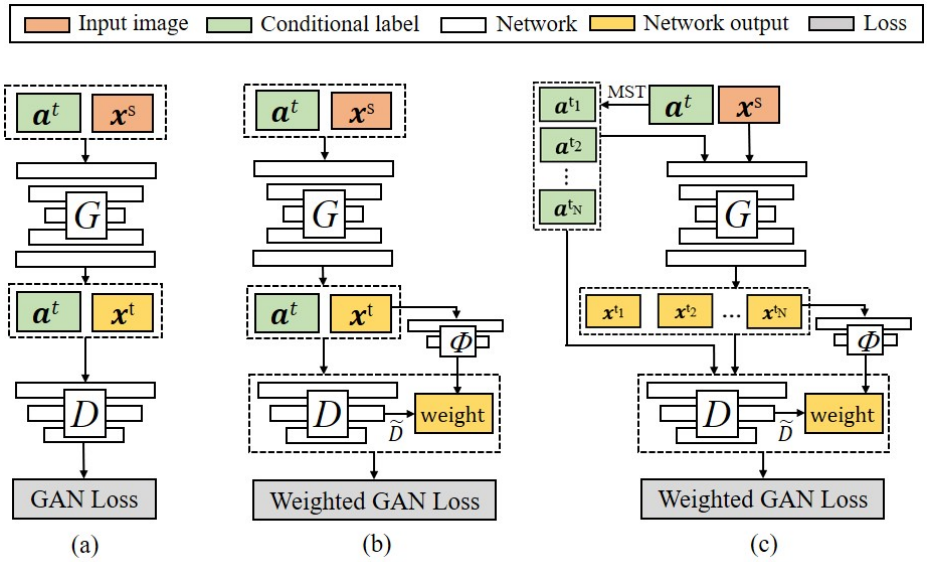


Fig. 2: A illustration about our framework. (a) is our backbone model, which is a Conditional GAN. (b) is the improved version with the proposed AIW (see Section 3.1). Our full model, which combines AIW and MST (see Section 3.2), is represented as (c).

3 Proposed Method

We consider visual domains characterized by an n-dimensional binary attribute vector $\boldsymbol{a} = [a_1, a_2, \cdots, a_n]^T$, where each bit a_i represents a meaningful visual attribute. We build Informative Sample Mining Network to learn all the mappings between these domains from unpaired training data. Our network takes a source image $\boldsymbol{x}^s$ with a target attribute $\boldsymbol{a}^t$ to produce the corresponding fake target image $\boldsymbol{x}^t$.

We adopt a Conditional GAN [24] as our backbone model, which is illustrated in Fig. 2 (a). That is, we train the generator and make the generator distribution p_g to capture the true data distribution p_{data}. Meanwhile, a discriminator tries to distinguish the real data from the synthesized fake data. The generator and the discriminator are trained jointly by optimizing the adversarial loss, which can be written as:

$$\min_G \max_D \mathcal{L} = \overbrace{\mathbb{E}_{\boldsymbol{x}^s, \boldsymbol{a}^s \sim p_{data}} [\log D(\boldsymbol{x}^s, \boldsymbol{a}^s)]}^{\mathcal{L}_{data}} + \overbrace{\mathbb{E}_{\boldsymbol{x}^t, \boldsymbol{a}^t \sim p_g} [\log(1 - D(\boldsymbol{x}^t, \boldsymbol{a}^t))]}^{\mathcal{L}_g}, \quad (2)$$

where $\boldsymbol{x}^t = G(\boldsymbol{x}^s, \boldsymbol{a}^t)$. For brevity, We use G and D to denote the generator and the discriminator, respectively. The inputs of G and D are the concatenation of an image and an attribute vector. We apply spatial replication on the attribute vector, making the sizes of the image and the attribute vector matched.

3.1 Adversarial Importance Weighting

In this section, we describe how to improve the sampling strategy of our backbone model, proposing Adversarial Importance Weighting. We will first consider the situation where we have the global optimal discriminator and then discuss the generalized situation.

To emphasize the contributions of informative samples, we improve the estimation of $\mathcal{L}_g$ by introducing an importance weight for each fake sample $\boldsymbol{x}^t$. Specifically, we aim to calculate the weight $\frac{p_{data}(\boldsymbol{x}^t)}{p_g(\boldsymbol{x}^t)}$, which is introduced in Eq. 1. To this end, we need to find a solution to make the weight computable. Recall the proposition made by Goodfellow et al. [7]: for any fixed G and any sample point $\boldsymbol{x}^t$, we have

$$D^*(\boldsymbol{x}^t) = \frac{p_{data}(\boldsymbol{x}^t)}{p_{data}(\boldsymbol{x}^t) + p_g(\boldsymbol{x}^t)}, \tag{3}$$

where D^* is the global optimal discriminator. To reveal the relation between the discriminator and the importance weight, let $D^*(\boldsymbol{x}^t) = S(\widetilde{D}^*(\boldsymbol{x}^t))$, where S denotes the sigmoid function. That is, we have

$$D^*(\boldsymbol{x}^t) = \frac{1}{1 + e^{-\widetilde{D}^*(\boldsymbol{x}^t)}}. \tag{4}$$

Combining Eq. 3 and Eq. 4, we can derive that $\frac{p_{data}(\boldsymbol{x}^t)}{p_g(\boldsymbol{x}^t)} = e^{\widetilde{D}^*(\boldsymbol{x}^t)}$. Hence, the weighted $\mathcal{L}_g$ can be formulated as:

$$\mathcal{L}_g = \mathbb{E}_{\boldsymbol{x}^t, \boldsymbol{a}^t \sim p_g}[e^{\widetilde{D}^*(\boldsymbol{x}^t)} \cdot \log(1 - D(\boldsymbol{x}^t, \boldsymbol{a}^t))]. \tag{5}$$

Eq. 5 indicates that a greater $\widetilde{D}^*(\boldsymbol{x}^t)$ brings $\boldsymbol{x}^t$ a bigger sample weight, which means $\boldsymbol{x}^t$ is more informative. In the meantime, a greater $\widetilde{D}^*(\boldsymbol{x}^t)$ also indicates $\boldsymbol{x}^t$ is harder to distinguish for the discriminator. Hence, similar to existing sample selection strategies [29,35,40], **mining the informative samples in GAN means finding the hard fake samples**.

Now we consider the practical situation, where we cannot get the optimal discriminator. A straightforward way to sidestep the need of D^* is replacing it with D. However, D may not provide accurate estimation if it is too far away from the optimality. Hence, we aim at measuring how close D is to D^*. Inspired by the fact [7] that D is close to D^* when p_g is similar to p_{data}, we propose a heuristic metric. Concretely, we first project each training batch $\{\boldsymbol{x}_1^s, \boldsymbol{x}_2^s, \cdots, \boldsymbol{x}_n^s\}$ and the corresponding generated results $\{\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_n^t\}$ onto a hypersphere whose radius is r by a pre-trained embedding model ϕ. Then, we can calculate the distance matrix $\boldsymbol{E}$, where $e_{ij} = ||\phi(\boldsymbol{x}_i^s) - \phi(\boldsymbol{x}_j^t)||$, i.e., the Euclidean distance between $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_j^t$ in the embedding space. We define that

$$\Delta l = \overbrace{(\sum_{i,j} e_{ij} - \sum_{i} e_i)}^{\Delta l_n} - \overbrace{\sum_{i} e_i}^{\Delta l_p} = \sum_{i,j} e_{ij} - 2 \cdot \mathrm{trace}(\boldsymbol{E}), \tag{6}$$

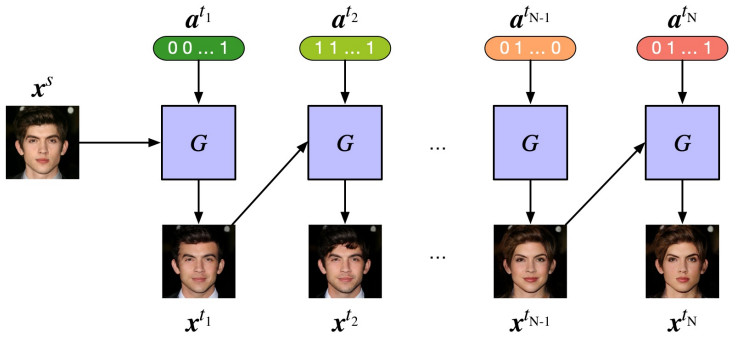


Fig. 3: An illustration about N-hop target image generation. After N times of translations, the generator produces the N-hop target image $\boldsymbol{x}^{t_N}$.

where Δl_p denotes the sum of the distances between the relative image pairs (e.g., e_{11}, e_{33}), and Δl_n denotes the sum of the distances between the permuted image pairs (e.g., e_{12}, e_{31}). Since only the source and the generated images that form a relative pair have the same content information, Δl_p should be as small as possible, and Δl_n should be as large as possible. In the optimal situation, we have $\Delta l_n^* = 2r$ and $\Delta l_p^* = 0$. Hence, Δl^* is a determined constant namely $2r$. In practical conditions, we can calculate $(\Delta l^* - \Delta l)$ to measure how close our network is to the global optimality. Formally, introducing the resale factor $(\Delta l^* - \Delta l)$ into the importance weight, we propose Adversarial Importance Weighting, which can be formulated as:

$$\mathrm{AIW}(\boldsymbol{x}^t) = \|1 + (\Delta l^* - \Delta l)\|^2 \cdot e^{\widetilde{D}(\boldsymbol{x}^t)}. \tag{7}$$

We have introduced AIW into our backbone model, as depicted in Fig. 2 (b). Accordingly, the formula of $\mathcal{L}_g$ is updated to:

$$\mathcal{L}_g = \mathbb{E}_{\boldsymbol{x}^t, \boldsymbol{a}^t \sim p_g}[\mathrm{AIW}(\boldsymbol{x}^t) \cdot \log(1 - D(\boldsymbol{x}^t, \boldsymbol{a}^t))]. \tag{8}$$

3.2 Multi-hop Sample Training

The proposed AIW makes the discriminator aware of hard samples and thus strengthens its power. However, since D and G are rivals during training, the training of G may become problematic due to the superior D. To address this issue, we introduce Multi-hop Sample Training which reduces sample hardness for the generator in a divide-and-conquer manner.

Let "hop" denote translation time a model takes to produce the target result. Fig. 3 provides a visual illustration, and here we give a formal definition:

$$\boldsymbol{x}^{t_N} = \overbrace{G(\ \cdots\ G(G}^{\mathrm{N}}(\boldsymbol{x}^s, \boldsymbol{a}^{t_1}),\ \boldsymbol{a}^{t_2}),\ \ \cdots),\ \boldsymbol{a}^{t_N}), \tag{9}$$

Algorithm 1: Training algorithm of INIT

Pretrain the embedding model ϕ
Initialize the generator G and the discriminator D
for *the number of training epochs* **do**
  Darw a training sample batch
  G forward propagates, producing 1-hop target images
  D forward propagates
  Calculate sample importance weights by Eq. 7
  Draw intermediate attributes
  G forward propagates, producing multi-hop target images
  D forward propagates
  Calculate $\mathcal{L}_{data} = \mathbb{E}_{\boldsymbol{x}^s, \boldsymbol{a}^s \sim p_{data}} [\log D(\boldsymbol{x}^s, \boldsymbol{a}^s)]$
  Calculate $\mathcal{L}_g$ by Eq. 10
  Calculate $\mathcal{L} = \mathcal{L}_{data} + \mathcal{L}_g$
  Optimize G by minimizing $\mathcal{L}$
  Optimize D by maximizing $\mathcal{L}$
end

where generator transforms $\boldsymbol{x}^s$ into $\boldsymbol{x}^{t_N}$ via N separate steps, and $\boldsymbol{x}^{t_N}$ is denoted as N-hop target image. We define $\{\boldsymbol{a}^{t_1}, \boldsymbol{a}^{t_2}, \cdots, \boldsymbol{a}^{t_{N-1}}\}$ as intermediate attributes, and $\{\boldsymbol{x}^{t_1}, \boldsymbol{x}^{t_2}, \cdots, \boldsymbol{x}^{t_{N-1}}\}$ are inferred to as intermediate images.

Previous approaches only consider the situation where $N = 1$. Consequently, some complex transformations may be too hard to learn for the generator. However, any complex transformation can be shrunk step-by-step (e.g., transfer multiple target attributes one-by-one), and it suffices to construct a feasible solution step-by-step. Therefore, we propose to reduce sample hardness by generating the target image in a multi-hop manner. Concretely, we introduce Multi-hop Sample Training, which considers {1-hop, 2-hop, $\cdots$, N-hop} target images. During training, we calculate losses on both the target images and the intermediate images, providing supervision information for each single step in the multi-hop image generation. Equipping the backbone model with MST, we update the formula of $\mathcal{L}_g$ to:

$$\mathcal{L}_g = \sum_{n=1}^{N} \mathbb{E}_{p_g}[\mathrm{AIW}(\boldsymbol{x}^t) \cdot \mathbb{E}_{p_{\text{n-hop}}}[\sum_{i=1}^{n} \log(1 - D(\boldsymbol{x}^{t_i}, \boldsymbol{a}^{t_i}))]], \tag{10}$$

where $\{\boldsymbol{a}^{t_1}, \boldsymbol{a}^{t_2}, \cdots, \boldsymbol{a}^{t_{n-1}}\} \sim p_{\text{n-hop}}$ $(n = 1, 2, \cdots, N)$, and we draw these intermediate attributes randomly in our experiments. Note that we also add AIW, which is proposed in Eq. 7, to this equation.

3.3 Implementation Details

Combining AIW and MST, our full model is able to select informative samples and then train them effectively. The complete diagram is depicted in Fig. 2 (c). During training, we follow the methodology of classical GAN [7] and optimize

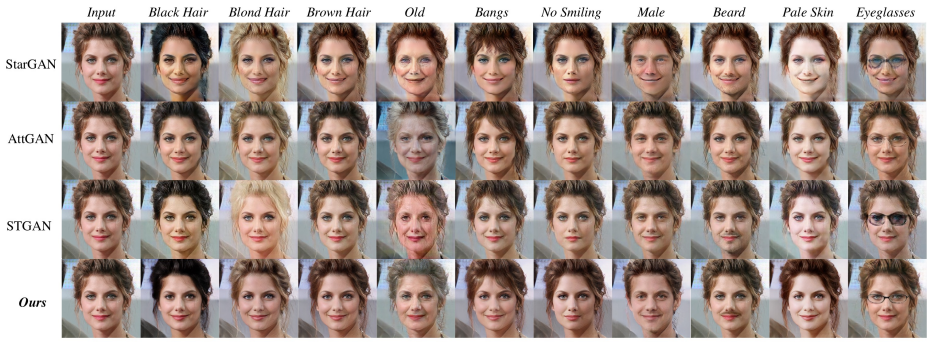


Fig. 4: Visual examples of single facial attribute editing results. From top to bottom, the rows are results of StarGAN [2], AttGAN [11], STGAN [21], and our INIT.

G and D iteratively. We first produce 1-hop target images, just like the previous approaches. Then, we estimate the importance weight by Eq. 7. Next, we draw intermediate attributes and produce multi-hop target images. Finally, we update model parameters by optimizing the weighted adversarial loss on multi-hop images. The training process is summarized in Algorithm 1.

In our experiments, we adopt 2-hop MST. We build a fully convolutional network as our generator and use a patch discriminator similar to [41]. The pre-trained VGG [31] is employed as the embedding model. Specifically, we use VGGFace [27] for facial attribute transfer. We optimize model parameters by Adam optimizer [17] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 1e-4. During testing, our generator directly produces the 1-hop target images as the output, which is the same as existing I2I methods.

4 Experiments

To validate the effectiveness of our approach, we perform extensive experiments on facial attribute transfer, season transfer, and edge&photo transfer. We produce 256×256 results and train our model as well as other competing models for 100 epochs. Our batch size is set to 16. In the following part, we first describe the datasets in our experiments (Section 4.1). Then we make comparisons with existing methods and report experimental results (Section 4.2, 4.3, and 4.4). Finally, we present an ablation study (Section 4.5).

4.1 Dataset

CelebA [23] is the largest publicly available dataset for multi-domain I2I tasks at present. There are annotations of 40 binary attributes for each image. In our experiment, we use the high-quality version, CelebA-HQ [16], for facial attribute transfer. We choose the following 10 attributes to construct the attribute

	StarGAN [2]	AttGAN [11]	STGAN [21]	Ours	Real Data
Hair Color	91.02	93.10	92.45	**94.47**	96.12
Aging	92.38	95.41	95.22	**97.90**	98.42
Bangs	87.97	91.03	91.84	**93.26**	93.67
Smile	85.53	90.47	87.41	**90.94**	91.00
Gender	90.72	**96.77**	94.76	96.57	98.25
Beard	87.90	93.53	93.09	**95.58**	95.34
Skin Color	89.35	92.65	94.03	**94.69**	94.22
Eyeglasses	93.38	96.44	96.09	**98.46**	99.31
FID	19.28	13.62	15.94	**11.16**	-

Table 1: The comparisons of the classification accuracy (%) [36] of each attribute (higher is better) and Fréchet Inception Distance [12] (FID, lower is better) on facial attribute transfer.

vector: *Black_Hair*, *Blond_Hair*, *Brown_Hair*, *Bangs*, *Smiling*, *Male*, *No_Beard*, *Pale_Skin*, and *Eyeglasses*. We randomly select 300 images as the testing set and use all the remaining images for training.

Yosemite Flickr Dataset [41] consists of 1,200 winter photos and 1,540 summer photos of Yosemite National Park. It is widely used for season transfer, which is an unpaired I2I problem. In our experiment, we follow the training and testing data divisions of CycleGAN [41].

Edge2Photo Dataset [5] contains photos of 250 categories of objects and the corresponding edges. Pix2pix [15] first uses the shoes category for edge-to-photo transfer. We follow the training and testing data divisions in Pix2pix to train our model.

4.2 Facial Attribute Transfer

We compare with StarGAN [2], AttGAN [11], and STGAN [21] on facial attribute transfer. We reproduce their results by the released source codes. The training and testing data for all the methods are the same.

Fig. 4 shows single attribute transfer results. Given an input image, each method produces 10 transformed images. For each output result, one specific attribute is toggled. For the relatively easy tasks like changing hair color, all the methods produce plausible results. By contrast, when dealing with the challenging ones like aging, our method produces more realistic results. In general, our INIT can achieve effective attribute transfer and outperform other methods.

We argue that **multiple attribute transfer** should also be emphasized. Transferring multiple attributes is at least no easier than transferring a single attribute, and the number of possible combinations is significantly larger. Hence, the hard cases are mainly from the multiple attribute conditions. We report comparison results of multiple facial attribute transfer in Fig. 1 and Fig. 5.

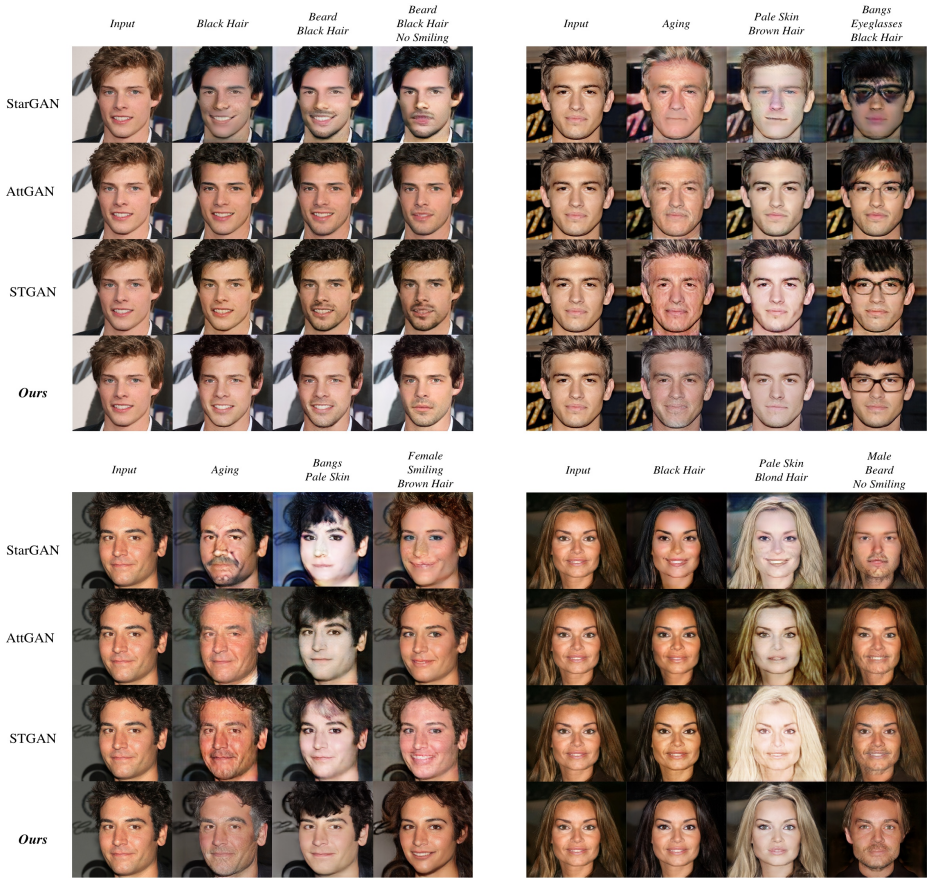
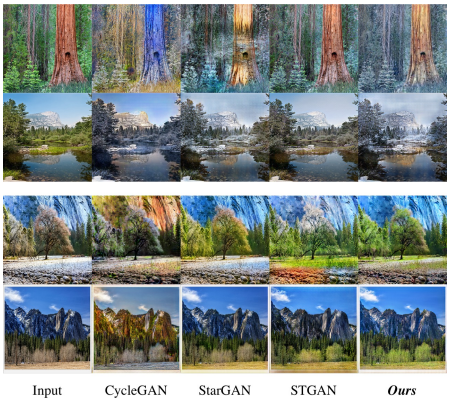


Fig. 5: Visual examples of multiple facial attribute editing, which is a very challenging case in multi-domain I2I. Please zoom in for better visualization. We make comparisons with StarGAN [2], AttGAN [11], STGAN [21].

In these cases, keeping visual quality and achieving effective attribute transfer become far more challenging. However, our INIT can still produce the desired results. Thanks to the weighting strategies, our method pays more attention to the hard cases and therefore yields better performance. Our superiority provides strong evidence on the effectiveness of sample selection.

We also calculate Fréchet Inception Distance (FID) [12] and the classification accuracy [36] to make objective comparisons. Lower FID is better since it means that the Wasserstein distance between the real distribution and the generated distribution is smaller. The classification accuracy reflects the effectiveness of attribute transfer, and thus higher is better. Following [21] , we train a Resnet-18 [10] as the classifier and calculate the accuracy on the transformed results. To train this classifier, we use the same data division of CelebA-HQ [16] as



(a) The two rows on the top are *summer→winter*, and the two rows on the bottom are *winter→summer*. We make comparisons with CycleGAN [41], StarGAN [2], and STGAN [21].

(b) The top two rows are *edge→shoes*, and the bottom two rows are *shoes→edge*. We compare with Pix2pix [15], StarGAN [2], and AttGAN [11]. The ground truths are on the upper left corner of the inputs.

Fig. 6: Visual examples of (a) season transfer and (b) edge&photo transfer.

the division for our generation tasks. In Table 1, we summarize FID and the classification accuracy of each class. It can be observed that our method has the best performance, indicating our improvements in visual realism and attribute transfer.

4.3 Season Transfer

For season transfer, we make comparisons with CycleGAN [41], StarGAN [2], and STGAN [21]. Note that only CycleGAN trains a pair of networks to achieve *summer→winter* and *winter→summer*, respectively. The other approaches can achieve season transfer by a single model.

We summarize the visual examples of translation results in Fig. 6a. We also calculate FID as an objective metric. Since real summer and winter photos have an apparent perceptual discrepancy, we calculate FID on the two seasons separately. Furthermore, we conduct a user study. We invite volunteers to select the best result among the transformed images from the four methods. All the testing images are compared, and we report the percent of votes for each method. The quantitative comparison results are summarized in the upper part of Table 2. The comparison on FID indicates that our approach favorably outperforms the competing methods, and we obtain the majority of the user votes.

4.4 Edge&Photo Transfer

In this subsection, our method is compared with Pix2pix [15], StarGAN [2], and AttGAN [11] on edge&photo transfer. Pix2pix needs to learn edge-to-photo

(a) **Season Transfer**				
Metric	CycleGAN [41]	StarGAN [2]	STGAN [21]	Ours
FID	48.40/49.39	44.84/52.16	45.61/50.37	**40.09/44.91**
Vote Percent	3.73%	11.07%	21.52%	**63.68%**
(b) **Edge&Photo Transfer**				
Metric	Pix2pix [15]	StarGAN [2]	AttGAN [11]	Ours
FID	39.59/17.13	33.23/13.92	29.01/13.86	**28.44/12.30**
Vote Percent	30.26%	3.77%	6.44%	**59.53%**

Table 2: The comparisons of FID and the percent of user votes (higher is better). For (a) season transfer, the FID is reported as "summer/winter". For (b) edge&photo transfer, the FID is reported as "photo/edge".

Model	AIW	MST	Hop Number	FID	Mean Acc
(a)	w/o	$w/$	2	18.91	92.86
(b)	$w/$	$w/$	3	11.23	94.98
(c)	$w/$	$w/$	2	**11.16**	**95.08**
(d)	$w/$	w/o	1	14.52	93.78
(e)	w/o	w/o	1	21.38	91.44

Table 3: Comparison results of different variations of our method. $w/$ and w/o are the abbreviations of "with" and "without", respectively. Our full model is equivalent to the variation (c).

and photo-to-edge separately, and the other methods can deal with the two translations simultaneously. Since we have paired data in this dataset, we add the L1 distance loss [15] in the pixel space, which is useful for paired I2I tasks. Note that we also add the L1 loss for the other competing methods to make fair comparisons.

We report the examples of translation results and the ground truth in Fig. 6b. Similar to season transfer, we calculate FID and conduct user study, the results of which are reported in the lower part of Table 2. Learning edge-to-photo and photo-to-edge as two separate tasks brings Pix2pix obvious advantages, but our method still has the best performance. Compared with StarGAN and AttGAN, our method produces more plausible results, showing stronger generalization ability for paired I2I tasks.

4.5 Ablation Study

In this subsection, we conduct an ablation study to verify the effectiveness of AIW and MST. To this end, we implement several variations of our approach and

evaluate them on facial attribute transfer. Concretely, we consider the following variations: (a) INIT without any importance sampling schemes, (b-d) INIT with n-hop sample training, where $n = 3, 2, 1$, respectively. (e) INIT that removes both AIW and MST, i.e., simply a conditional GAN [24]. Note that our full model is equivalent to variation (c).

We use the same experiment setting and train these variations for the same number of iterations. Note that variations with a smaller hop number will have more iterations for training new samples since they have fewer intermediate results to optimize. Tabel 3 shows comparison results on quantitative metrics, and please refer to our Supplementary for visual examples. Through the ablation study, we can verify the following two points:

Mining informative samples plays an important role. Without the important sampling scheme, the performances of variations (a) and (e) drop sharply. Even when we double the training iterations of variation (e), its performance is still obviously inferior. Hence, merely taking more training time is not an effective option.

The optimal choice is 2-hop sample training. As the hop number increases, more intermediate samples are drawn during training. It means that we pay more attention to reduce the sample hardness for the generator. However, as indicated by Eq. 10, it also means 1-hop samples contribute less to the loss function. Note that during testing, the evaluations are based on 1-hop target images. Hence, a larger hop number does not guarantee better performance in practice.

5 Conclusion

In this paper, we propose to integrate Importance Sampling into a GAN-based model, resulting in Adversarial Importance Weighting for high-quality multi-domain image-to-image translation. Furthermore, Multi-hop Sample Training subtly reduces sample hardness while preserving sample informativeness. Thanks to the improvements in training efficiency, our approach achieves effective translation even when dealing with a large number of challenging visual domains. We conduct extensive experiments on practical tasks, including facial attribute transfer, season transfer, and edge&photo transfer. The results consistently demonstrate our superiority over existing methods.

Acknowledgement. This work is funded by the National Natural Science Foundation of China (Grant No. U1836217), Beijing Natural Science Foundation (Grant No. JQ18017) and Youth Innovation Promotion Association CAS (Grant No. Y201929).

References

1. Cao, D., Zhu, X., Huang, X., Guo, J., Lei, Z.: Domain balancing: Face recognition on long-tailed domains. In: CVPR (2020)
2. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
3. Deng, Q., Cao, J., Liu, Y., Chai, Z., Li, Q., Sun, Z.: Reference guided face component editing (2020)
4. Duan, Y., Zheng, W., Lin, X., Lu, J., Zhou, J.: Deep adversarial metric learning. In: CVPR (2018)
5. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? In: SIGGRAPH (2012)
6. Evans, M., Swartz, T., et al.: Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. Stat Sci (1995)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
8. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: CVPR (2020)
9. Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T., et al.: Smart mining for deep metric learning. In: ICCV (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: Facial attribute editing by only changing what you want. TIP (2019)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a Nash equilibrium. In: NeurIPS (2017)
13. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: CVPR (2014)
14. Huang, C., Loy, C.C., Tang, X.: Local similarity-aware deep feature embedding. In: NeurIPS (2016)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
19. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. In: NeurIPS (2017)
20. Law, M.T., Thome, N., Cord, M.: Quadruplet-wise image similarity learning. In: ICCV (2013)
21. Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., Wen, S.: STGAN: A unified selective transfer network for arbitrary image attribute editing. In: CVPR (2019)
22. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017)
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)

24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
25. Oh, M.S., Berger, J.O.: Integration of multimodal functions by monte carlo importance sampling. J AM STAT ASSOC (1993)
26. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: ICML (2016)
27. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC (2015)
28. Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. Signal Process. Mag. (2015)
29. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
30. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: ICCV (2015)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Veach, E., Guibas, L.J.: Optimally combining sampling techniques for monte carlo rendering. In: SIGGRAPH (1995)
33. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: ICCV (2017)
34. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR (2014)
35. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: ICCV (2017)
36. Wu, P.W., Lin, Y.J., Chang, C.H., Chang, E.Y., Liao, S.W.: RelGAN: Multi-domain image-to-image translation via relative attributes. In: ICCV (2019)
37. Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., Bai, X.: Hard-aware point-to-set deep metric for person re-identification. In: ECCV (2018)
38. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: ICCV (2017)
39. Zhao, Y., Jin, Z., Qi, G.j., Lu, H., Hua, X.s.: An adversarial approach to hard triplet generation. In: ECCV (2018)
40. Zheng, W., Chen, Z., Lu, J., Zhou, J.: Hardness-aware deep metric learning. In: CVPR (2019)
41. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)